Apprentissage adverse et algorithmes de détection d'évènements : une première typologie

Olivier de Casanove
Institut de Recherche en Informatique de Toulouse - IRIT
Université Toulouse III - Paul Sabatier
Toulouse, France
olivier.decasanove@irit.fr

Florence Sèdes
Institut de Recherche en Informatique de Toulouse - IRIT
Université Toulouse III - Paul Sabatier
Toulouse, France
florence.sedes@irit.fr

***Abstract*—Les réseaux sociaux peuvent être un support opérationnel pour la détection d'évènements. La littérature se concentre sur l'amélioration de la performance des algorithmes de détection d'évènements, mais fait peu mention de mesures permettant de se défendre contre une potentielle manipulation de la donnée d'entrée (appelé "adversarial learning" ou encore "apprentissage adverse"). Toute personne le souhaitant peut poster des messages sur les réseaux sociaux, il est donc possible pour n'importe qui de manipuler les données utilisées pour la détection. En conséquence, les algorithmes de détection d'évènements actuels sont particulièrement vulnérables à l'apprentissage adverse. Cet article propose une première typologie pour décrire les menaces liées à l'apprentissage adverse contre les algorithmes de détection d'évènements.**

***Index Terms*—Apprentissage Adverse, Détection d'évènements, Réseaux sociaux, Sécurité**

I. Introduction

Les plateformes de microblogs permettent aux utilisateurs d'échanger des contenus numériques de taille réduite. De ces données numériques publiées en temps réel peuvent être extraites des informations sur des évènements. Un évènement, dans le contexte des plateformes de microblogs et des réseaux sociaux, peut être compris comme un fait marquant se déroulant à un instant t, en dehors de l'espace numérique et qui génère une discussion associée à ce sujet par les utilisateurs.

Pour détecter ces évènements dans le flux de données émis par les réseaux sociaux, nous utilisons des algorithmes qui prennent en entrée ledit flux et génèrent des clusters de messages, où chaque cluster caractérise un évènement. Dans la revue de la littérature [1], les auteurs dressent une liste des applications de ces algorithmes.

Indépendamment de la technique utilisée par l'algorithme de détection d'évènements, la littérature se concentre surtout sur l'amélioration des performances de leur détection. La question de la sécurité de ces algorithmes n'est pas abordée ; sur des plateformes qui intègrent des messages comme données d'entrée, elle doit être posée. Si l'hypothèse implicite : "les données ne sont pas polluées par des messages intentionnellement formés pour nuire à la détection d'évènements" peut avoir un sens dans des contextes spécifiques, elle se révèle fausse dans d'autres. Par exemple, lors de la détection d'évènements relatifs à la sécurité informatique ou à une quelconque forme de surveillance, il faudra lutter face à des personnes malintentionnées ou des bots. On peut pourtant facilement trouver des papiers dans ces domaines [2] [5], mais qui ne tiennent pas compte de potentiels acteurs malveillants. Lorsque cette hypothèse est considérée comme fausse, nous nous situons dans un contexte d'apprentissage adverse (*"adversarial learning"* en anglais). L'hypothèse avec laquelle nous travaillons devient alors : "Les données d'entrée sont corrompues par un acteur malveillant dans le but de nuire au bon fonctionnement de notre algorithme." L'objectif de ce papier est de présenter une première typologie permettant de modéliser ce contexte d'apprentissage adverse sur les réseaux sociaux. Dans la première partie, nous détaillerons les profils d'attaquants. Dans la deuxième partie, nous verrons les attaques possibles. Dans la troisième partie, nous discuterons brièvement de la suite de ce travail. En dernière partie, nous apporterons une conclusion.

II. Les attaquants

La typologie que nous proposons repose sur trois éléments : la cible, l'attaquant et le mode d'attaque. La cible a déjà été définie comme étant les algorithmes de détection d'évènements. Concernant l'attaquant, il n'existe pas de revue de la littérature pour lister tous les profils qui pourraient s'attaquer à des algorithmes analysant le flux de messages des réseaux sociaux. On retrouve cependant dans d'autres contextes la mention de profils particuliers qui peuvent nous aider à mieux définir l'attaquant. Nous allons détailler ces profils.

A. Le troll

Le troll est défini par [4] comme un acteur malveillant dont le but est "de propager des rumeurs, de désinformer, de spéculer et d'exploiter les biais cognitifs ainsi que les mauvais comportements pour manipuler l'opinion des utilisateurs". Bien souvent les trolls ne sont que des programmes qui sont opérés par un utilisateur final. La cible des trolls étant les utilisateurs des réseaux sociaux, notre algorithme de détection d'évènements n'en subira pas d'attaque directe. Cependant, à cause de leur désinformation, ils dégradent la qualité de l'information disponible dans le flux de messages.

B. Les spammeurs

Un spammeur est un utilisateur, humain ou robot, qui crée des tweets malintentionnés, polluant ainsi le flux de messages des réseaux sociaux. Le spam peut avoir plusieurs buts, le plus évident étant la promotion d'un contenu malveillant permettant ensuite un gain monétaire. Grâce à un jeu de données regroupant de nombreux spams, [3] a extrait quatre profils de spammeurs :

- Spammeur de doublons : il reposte en permanence le même message ou des messages avec très peu de variations.
- Spammeur de doublons @ : il possède un comportement identique au précédent, mais utilise les mentions pour interpeler des utilisateurs.
- Promoteur malveillant: il poste un mélange de messages légitimes et de spams pour promouvoir, par exemple, un produit ou une idée. Dans sa définition, il se rapproche du troll.
- Infiltrateur d'amis : il apparait comme un utilisateur légitime, mais abuse de la confiance liée aux relations entre utilisateurs sur le réseau social.

Les profils que nous venons de définir s'attaquent principalement aux utilisateurs. L'algorithme de détection d'évènements n'ayant ni ami ni capacité à faire une action permettant une quelconque rétribution des spammeurs, il n'en est pas leur cible. Cependant l'envoi massif de messages pollue la donnée d'entrée et rend la détection d'évènements plus ardue.

C. Les adversaires

En créant un algorithme de détection d'évènements spécifiques aux CVEs (Common Vulnerability Exposure), [6] identifie trois types d'acteurs qui pourraient attaquer leur algorithme :

- Adversaire bavard : il n'est pas au courant de la technologie utilisée, il envoie des messages composés de données aléatoires.
- Adversaire imitateur de mots : il connait la technologie utilisée ainsi que les notions de base liées à la détection d'évènements. Ces ressources sont limitées et ne peuvent pas manipuler toutes les statistiques fournies . Il ne peut pas, par exemple, posséder de compte sur les réseaux sociaux qui soit crédible.
- Adversaire imitateur complet : il connait la technologie utilisée et a une maitrise du sujet. Ses ressources ne sont pas contraintes et il peut posséder de nombreux comptes crédibles.

Parmi les profils attaquants précédemment listés, les trois profils de cette sous-section sont particulièrement pertinents au regard de notre sujet d'étude. À notre connaissance, c'est d'ailleurs le seul travail où il est fait référence à de potentiels adversaires dans la détection des évènements sur les réseaux sociaux.

D. Résumé

Des précédents profils énoncés, nous extrayons la classification suivante, avec une définition relative aux impacts de l'attaquant sur les algorithmes de détection d'évènements :

- Troll : son objectif est de faire disparaitre un sujet ou faire apparaitre de faux sujets. Il n'attaque pas directement l'algorithme de détection d'évènements, mais sa pollution nuit à la qualité de l'information présente dans les messages.
- Spammeur : il crée un grand nombre de messages dans un but qui lui est propre. Il n'attaque pas directement l'algorithme de détection d'évènements, mais à cause de lui la qualité du flux d'information se trouve dégradée. De plus, un trop grand nombre de messages à traiter peut provoquer des problèmes de performance dans l'algorithme de détection d'évènements.
- Adversaire : son objectif est spécifiquement de nuire à l'algorithme de détection d'évènements. Ses moyens peuvent être divers et variés et on peut lui supposer une connaissance au moins basique des technologies de détection d'évènements.

Dans un second temps, ces profils attaquants peuvent être ensuite distingués sur deux autres critères. Ils peuvent être sachants ou ignorants du système de détection utilisé. Enfin les attaquants peuvent être libres ou contraints par un certain cadre. Il faut comprendre par là, qu'ils puissent être soumis à des contraintes légales sur leurs actions ou bien simplement avoir des moyens financiers variables. Maintenant que nous connaissons la cible de l'attaque, les algorithmes de détection d'évènements, et les profils des attaquants, intéressons-nous aux attaques possibles.

III. Les attaques

En sécurité informatique il est courant d'utiliser le modèle CIA, pour Confidentialité, Intégrité et Disponibilité (Availability en anglais). Cependant ce modèle ne permet pas de représenter avec précision les problématiques liées à l'apprentissage adverse dans le contexte de la détection d'évènements sur les réseaux sociaux. On peut par exemple argumenter sur l'utilité de la notion de confidentialité, lorsque toutes les données d'entrée sont publiques. À la place nous utiliserons la notion de validité et de fiabilité.

A. Validité

Le concept de validité désigne la capacité d'un test à détecter ce qu'il prétend détecter. Dans notre cas, est-ce que notre détecteur d'évènements réussit réellement à détecter de vrais évènements ? Du point de vue de l'attaquant, l'objectif est de créer le plus de faux évènements possibles et de réduire au silence le plus de vrais évènements.

B. Fiabilité

Le concept de fiabilité désigne la capacité d'un test à être consistant, c'est-à-dire qu'une même donnée d'entrée donnera toujours une même donnée de sortie. Adaptée à notre contexte d'apprentissage adverse, la question devient : est-ce qu'un

évènement, respectivement une absence d'évènement, sera toujours considéré comme tel si un acteur malveillant pollue les données d'entrée ? Une autre question plus difficile à formaliser est : un même évènement détecté, aura-t-il toujours la même proportion de messages significatifs si un acteur malveillant pollue les données d'entrée ?

C. Classification des attaques

Après avoir étudié les algorithmes de détection d'évènements sur les réseaux sociaux, nous avons pu identifier plusieurs attaques. Ces attaques peuvent être regroupées dans trois grandes catégories : *création, dispersion ou modification d'évènements*. Les catégories regroupent des scénarios d'attaque qui utilisent les mêmes moyens, mais n'ont pas le même but. Ces scénarios sont ensuite décrits en fonction de leur impact sur la validité et la fiabilité de la détection.

1) Création d'évènements : l'adversaire arrive à déclencher une détection d'évènement, ce qui augmente le nombre de faux positifs. Le moyen est simple : utiliser un outil pour générer procéduralement des messages. Les messages sont ensuite injectés dans le flux de données. Nous recensons trois scénarios différents entrant dans cette catégorie d'attaque :

- Fabrication : des messages malveillants déclenchent un évènement. Le nombre de faux positifs augmente.
- Extension de messages : de vrais messages ne faisant pas partie d'un évènement, associés à des messages malveillants, déclenchent un nouvel évènement. Un vrai négatif devient un faux positif. Le nombre de faux positifs augmente.
- Rejeu : un vrai évènement est rejoué par un attaquant. Le nombre de faux positifs augmente.

2) Dispersion d'évènements : l'adversaire arrive à détériorer la qualité de la détection de certains évènements, allant parfois jusqu'à ne pas faire détecter l'évènement du tout. Pour cela l'adversaire va générer un grand volume de messages qui seront injectés entre deux messages significatifs, l'objectif étant que les messages significatifs deviennent suffisamment distants pour ne plus être associés lors de la découverte d'évènements. Trois scénarios identifiés rentrent dans cette catégorie :

- Fragmentation : un évènement déclenche une alerte plusieurs fois. Un vrai positif en devient plusieurs créant une inconsistance.
- Annulation : un évènement ne déclenche pas d'alerte alors qu'il aurait dû. Un vrai positif devient un faux négatif. Le nombre de vrais positifs diminue.
- Détérioration : le nombre de messages pertinents associés à un évènement est diminué. Les évènements détectés sont moins significatifs.

3) Modification d'évènements : l'adversaire arrive à faire changer le sujet même de l'évènement pour l'algorithme de détection. Il va générer des messages procéduralement qui peuvent être associés à un évènement déjà présent, mais qui en changeront les caractéristiques. Deux scénarios appartiennent à cette catégorie :

- Dérive : l'adversaire arrive à faire changer le sujet de l'évènement en y injectant de faux messages. Les évènements détectés sont moins significatifs.
- Fusion : l'adversaire arrive à faire converger le sujet d'un évènement vers un autre, entrainant la fusion de deux évènements distincts.Le nombre de vrais positifs baisse.

IV. Discussion

À notre connaissance, peu de papiers sur le sujet prennent en compte ce contexte. Sabottke et. al tente de résoudre le problème en ajoutant un système de liste blanche [6]. Cet ajout leur permet d'obtenir de meilleurs résultats. Sceller et. al. propose un détecteur d'évènements de cybersécurité [7]. La sécurité de leurs données d'entrée est traitée comme un potentiel point d'amélioration. Un premier niveau de sécurité est la mise en place d'un filtre antispam, sujet déjà abordé au sein de l'équipe [8], ce qui a pour conséquence directe de compliquer la création de tweets malveillants. Nous souhaitons, dans le futur, utiliser un détecteur d'évènements sur les réseaux sociaux pour étudier les attaques informatiques. Nous avons donc dû travailler sur la sécurité de ces algorithmes, la première étape étant cette typologie pour mieux cerner la menace.

V. Conclusion

Dans ce papier nous avons mis en avant le faible niveau de recherche sur la sécurité des algorithmes de détection d'évènements, qui s'avère pourtant un point crucial compte tenu de l'usage croissant des réseaux sociaux. Nous avons proposé une typologie explicitant quels pouvaient être les profils d'attaquants et leurs attaques possibles.

References

[1] Atefeh F., Khreich W. (2015). A Survey of Techniques for Event Detection in Twitter. Computational Intelligence, vol. 31, no1, p. 132–164.

[2] Khandpur R. P., Ji T., Jan S., Wang G., Lu C.-T., Ramakrishnan N. (2017). Crowdsourcingcybersecurity: Cyber attack detection using social media. InProceedings of the 2017 ACMon Conference on Information and Knowledge Management, p. 1049–1057.

[3] Lee K., Eoff B. D., Caverlee J. (2011). Seven months with the devils: A long-term study ofcontent polluters on twitter. In Fifth international aaai conference on weblogs and social media.

[4] Mihaylov T., Georgiev G., Nakov P. (2015). Finding Opinion Manipulation Trolls in News Community Forums.

[5] Ritter A., Wright E., Casey W., Mitchell T. (2015). Weakly supervised extraction of computersecurity events from twitter. InProceedings of the 24th international conference on worldwide web, p. 896–905. Republic and Canton of Geneva, CHE, International World Wide Web Conferences Steering Committee.

[6] Sabottke C., Suciu O., Dumitras T. (2015). Vulnerability disclosure in the age of so-cial media: Exploiting twitter for predicting real-world exploits. In24th USENIX securitysymposium (USENIX security 15), p. 1041–1056. Washington, D.C., USENIX Association.

[7] Sceller Q. L., Karbab E. B., Debbabi M., Iqbal F. (2017). Sonar: Automatic detection of cybersecurity events over the twitter stream.Proceedings of the 12th International Conferenceon Availability, Reliability and Security.

[8] Washha M., Qaroush A., Mezghani M., Sedes F. (2017). A Topic-Based Hidden Markov Model for Real-Time Spam Tweets Filtering. In Procedia Computer Science, Volume 112, Pages 833-843